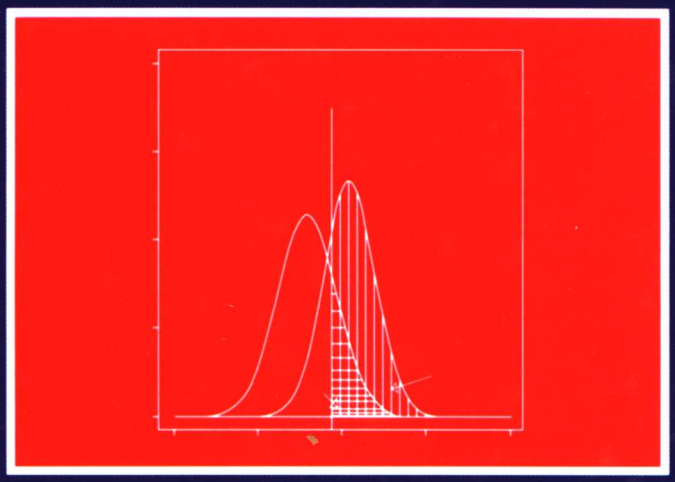
诊断医学统计学

Statistical Methods in Diagnostic Medicine



原著 Xiao-Hua Zhou
Nancy A. Obuchowski
Donna K. McClish
译者 宇传华

人民卫生出版社

诊断医学统计学

Statistical Methods in Diagnostic Medicine

原　著　Xiao-Hua Zhou
Nancy A. Obuchowski
Donna K. McClish

译　者　宇传华

人民卫生出版社

图书在版编目（CIP）数据

诊断医学统计学/宇传华译. —北京：
人民卫生出版社，2005.2
ISBN 7-117-06597-4

Ⅰ.诊… Ⅱ.宇… Ⅲ.实验室诊断-医学统计
Ⅳ.R446-32

中国版本图书馆 CIP 数据核字（2005）第 007037 号

图字：01-2004-4952

诊断医学统计学

译　　者：宇传华
出版发行：人民卫生出版社（中继线 67616688）
地　　址：（100078）北京市丰台区方庄芳群园 3 区 3 号楼
网　　址：http://www.pmph.com
E-mail：pmph @ pmph.com
印　　刷：北京市安泰印刷厂
经　　销：新华书店
开　　本：787×1092　1/16　　**印张：**19.25
字　　数：436 千字
版　　次：2005 年 3 月第 1 版　2005 年 3 月第 1 版第 1 次印刷
标准书号：ISBN 7-117-06597-4/R·6598
定　　价：49.00 元

前　言

估计与比较诊断试验准确度，在诊断医学研究中具有十分重要的意义。尽管诊断准确度方法学的研究取得了显著的进展，但迄今为止，还没有一本教科书可供参考；多年来，有许多临床和统计界的同事询问有关诊断试验准确度的问题，为此，我们撰写了这本书。书中全面考虑了诊断研究设计与分析的统计学方法，如样本含量的计算、诊断试验准确度的估计、竞争诊断试验准确度的比较、诊断准确度数据的回归分析等。此外，本书也讨论了一些最新发展的方法学论题，如证实偏倚与不完善参考偏倚的校正、群集诊断准确度数据的分析、Meta 分析等。

本书共分 12 个章节。前三章讨论了诊断准确度的常用指标与诊断准确度的研究设计，第 1 章给出了一些诊断准确度研究中的统计学问题；第 2 章定义了几个常用的诊断准确度指标，如灵敏度、特异度、预测价值、接受者工作特征（receiver operating characteristic，ROC）曲线及其有关指标；第 3 章描述了避免常见诊断准确度偏倚的临床研究设计。

第 4 和第 5 章讨论了诊断准确度的估计与假设检验方法。第 4 章描述了估计灵敏度、特异度、预测价值和 ROC 曲线的方法；第 5 章给出了比较几个竞争试验相对准确度的方法。第 6 章阐明了诊断准确度研究的样本含量计算；第 7 章讨论了诊断试验研究中 Meta 分析的非数学问题。

第 8 至第 12 章讨论了更高级的分析技术。第 8 章运用回归模型研究了患者特征对诊断试验准确度的影响；第 9 章探讨了比较几条相关 ROC 曲线（如多阅片者研究）的方法；第 10 章给出了校正证实偏倚的估计与推断方法；第 11 章讨论了采用不完善金标准时，诊断准确度的正确估计方法；第 12 章描述了诊断试验研究的 Meta 分析的统计学方法。

本书的读者对象是对诊断研究感兴趣的临床医生，对分析诊断研究数据感兴趣的统计工作者，以及有志于诊断医学统计学研究的统计学家与研究生。

读者应具备统计学方法的基本知识。阅读第 8、9 章的读者需要熟悉回归模型知识；熟悉缺失数据的统计学方法将有助于读者阅读第 10 和第 11 章。

译 者 序

在完成我的博士学位论文《ROC 分析方法及其在医学研究中的应用》的文献复习阶段，原作者周晓华教授在诊断试验评价方面所做的研究深深吸引了我；2001 年在中山大学做博士后研究期间，又有幸与周教授合作，并欣喜得知他及其同事即将出版《Statistical Methods in Diagnostic Medicine》一书的消息。一年的期待后，这本书终于问世了，于 2002 年 7 月，著名的 John wiley & Sons 公司正式出版了这本书。于 2002 年 12 月，周晓华教授趁回国讲学的机会，特地赠送给了我这本书。

潜心研读后，我发现这本书不仅内容丰富，具有较高的研究参考价值，而且也具有较强的实际应用价值，尤其对于我国的临床医生和医学统计学工作者。为此，我决定将这本书的英文版翻译成中文。

这本书有如下特点：

1. 循证医学的重要组成部分

临床医疗包括诊断和治疗两个方面，医生对疾病作出准确的诊断，毫无疑问对疾病的正确治疗有帮助。一个试验的诊断准确度有多大？多个试验中，哪一个试验的诊断准确度更高？如何进行诊断试验准确度的研究？回答这些问题涉及到很多统计学知识，这本书从诊断试验的研究设计到数据分析，以理论与实例相结合的方式，较为全面地介绍了诊断医学中所涉及的统计学方法。这些方法为寻找最佳的医学诊断证据、指导临床诊治决策提供了重要的工具。

2. 内容全面丰富

国际上涉及诊断试验研究设计与分析的综合性信息资源几乎没有，这本书正好满足了这方面的需要。全书共分为两大部分，第一部分（第一篇）涉及到诊断试验研究设计与分析的基础知识，适用于一般的临床医生；第二部分（第二篇）给出了诊断试验研究的更高级方法，适用于对诊断试验研究方法感兴趣的统计学工作者和其他研究人员。

3. 理论与实践相结合

本书在介绍了诊断试验研究的概念、公式等理论之后，紧接着以实际数据加以阐述，有利于读者理解与巩固所学的理论知识。书中的每一章末尾（除第 1 章外）附有大量的练习题，适用于读者检验自己的学习情况。每章后附有大量参考文献目录可供读者查阅参考。

第四军医大学郭秀娥副教授参与了本书第 9 章的翻译。华中科技大学同济医学院公共卫生学院张述林书记以及我的硕士导师余松林教授、解放军总医院姚晨教授、第一军医大学陈平雁教授、广州医学院王心旺博士、吕嘉春博士等对译者的整个翻译工作给予了大力的精神支持。原版书的第一作者、美国华盛顿大学的周晓华教授及时回答了译者的大量咨询问题，并提供了原版书的印刷错误勘误表。

本书的出版得到了国家自然科学基金（编号：30371254）和华中科技大学同济医学

院人才引进启动基金的资助。在此一并致谢!

由于译者水平有限，翻译过程中难免出现错误，欢迎读者采用如下通讯方式与译者联系，以便译者及时对错误加以更正。

宇传华

E-mail:yuchua@163.com

个人主页:http://vip.6to23.com/statdtedm

2004 年 6 月 16 日于武汉华中科技大学同济医学院

目　录

第一篇　基本概念和方法

第二篇　高 级 方 法

第1章 绪 论

1.1 写这本书的目的

诊断试验对医疗保健具有十分重要的作用，对卫生保健费用的影响也具有十分重要的意义（Epstein，Begg，and McNeil，1986）。然而，诊断试验研究的质量并不高（Begg，1987）。1995年Reid、Lachs和Feinstein查阅1978～1993年期间发表的有关诊断试验文献后，指出了诊断试验设计与分析的许多错误。这些错误导致了人们对诊断试验研究结果的不信任，对诊断试验的选择与解释产生了负面的影响。

下面列举三个诊断试验研究的常见错误。第一个常见错误是关于诊断试验的解释，许多新诊断试验研究者企图仅基于健康志愿者的试验结果，建立解释这些试验的标准。例如，胰腺炎新诊断试验研究者通过测量健康志愿者的某酶含量，以偏离该酶含量均数的三个标准差（SD）作为决策标准或诊断界点（cut-off point），如果患者该酶含量低于健康志愿者均数的3个SD则记为胰腺炎阳性，高于这个诊断界点则记为胰腺炎阴性。这一陈述的错误为：

①没有考虑数据的确切分布，即是否数据服从正态（高斯）分布；

②没有考虑健康自愿者与病人的试验结果在数量上的可能重叠程度；

③没有考虑诊断错误的临床意义。诊断错误有两种：一是无该病者被错误判为阳性（误诊），二是病人被错误判为阴性（漏诊）；

④对健康志愿者结果的不恰当推广。

在本书第2章，将讨论确定最佳诊断试验界点的有关影响因素，第4章还将深入介绍寻找最佳诊断试验界点，估计有关诊断错误的方法。

为了确定患者的真实疾病状况（有与无病），有时需要采用较苛刻的评价方法（如剖腹探查），第二种诊断试验研究的常见错误是关于对所采用苛刻评价的看法，这种错误的观点认为：科学研究允许将少量没有接受苛刻评价的患者忽略不计。例如，通气灌注肺扫描诊断肺栓塞是一种无创伤性试验，用于筛查肺栓塞高危人群，各种人群中的准确度未知；肺血管造影诊断肺栓塞具有较高的诊断准确度，但它对人体具有较大损害。在评价通气灌注肺扫描准确度的研究中，通常只将同时接受通气灌注肺扫描和肺血管造影的患者作为研究样本，并以血管造影作为估计准确度的参考（即金标准，其定义和实例见第2章），仅接受通气灌注肺扫描但未接受血管造影者将从研究中剔除。

上述研究设计可能导致严重的试验准确度估计错误，原因是研究样本并没有真实代表接受通气灌注肺扫描的患者总体，扫描结果阳性者通常被推荐做血管造影，而对于扫描结果阴性者，为了避免冒不必要的风险，通常不主张继续做血管造影。第3章将讨论

全面检查偏倚（workup bias）及其最常见形式——证实偏倚（verification bias），并讨论避免它们的措施。第10章将给出校正证实偏倚的统计学方法。

第三个常见错误与一致性问题有关，研究者通常错误地基于与常规试验的一致程度，下某新试验诊断能力的结论。例如，数字乳腺X线照相是一种新的胸部影像筛查与诊断方法，与常规胸片相比有许多优点，如影像容易存贮、传递等。在比较这两个试验的研究中，如果结果一致则可为该新试验感到欣慰。但如果数字与胶片结果不一致又会怎样呢？武断地下数字胸片准确度较差的结论显然不正确，如果数字X线照相与常规胸片相比有较好的准确度，这两个诊断试验也将不一致；类似地，两个诊断试验有相同的准确度，但如果被错误诊断患者各有所不同，也可能获得较差的一致性。评价某新试验诊断价值的更有效方法应该是：在已知真实诊断结果的情况下，估计与比较两个试验的准确度。诊断准确度评价比一致性评价更困难，但更合理、更有效（Zweig and Campbell，1993）。第5章将给出真实诊断结果已知情况下，比较两诊断试验准确度的方法；第11章还要给出真实诊断结果未知情况下，比较两诊断试验准确度的方法。

毫无疑问，诊断试验准确度研究是设计的挑战，其数据分析需要专门的统计学方法。目前，好的参考文献很少，涉及诊断试验研究设计与分析的综合性信息资源几乎没有，这本书正好满足了这方面的需要，给出并阐明了诊断试验准确度研究设计、分析、解释和报告的方法与概念。第Ⅰ篇（第2～7章）将定义几个诊断试验准确度的指标，描述设计诊断试验准确度研究的策略，给出估计与比较诊断试验准确度、计算样本含量和综合文献Meta分析的基本统计学方法。第Ⅱ篇（第8～12章）将给出试验准确度研究中更高级的统计学方法，如多阅片者研究的分析、具有证实偏倚或不完善金标准研究的分析、诊断试验的Meta分析等。

1.2 什么是诊断准确度

诊断试验有两个目的（Sox，Jr. et al.，1989）：①提供患者疾病的可靠信息，②影响医生的治疗患者计划。McNeil和Adelstein（1976）还补充了第三个可能的目的：通过研究，了解疾病的机制和自然病史（如对慢性病患者的重复试验）。只要医生懂得如何解释某试验，试验便可达到这些目的。上述信息通过试验的诊断准确度评价获得，简单地说，诊断准确度就是试验区分备选疾病状态的性能（Zweig and Campbell，1993），尽管可能的疾病状态有两种以上，但临床上通常可以将它们合理地划分成两类，如有、无帕金森病，有、无扩散性癌症等。本书只考虑疾病状态为二分类的情况。

评价某诊断试验的性能时，希望知道这两种疾病状态的试验结果有无差异，如果试验结果无差异，则该诊断试验准确度十分小，可以忽略不计；如果有差异且无重叠，则该诊断试验具有完善的准确度；大多数试验的准确度落在这两个极端之间。要避免的最主要错误是假定试验结果能真实反映患者的情况（Sox，Jr. et al.，1989），大多数诊断信息是不完善的，这可能影响医生的思想，但患者真实情况的不确定性依然存在。如果试验阴性，医生应该假定患者无此病，并且让他回家吗？如果试验阳性，医生应该假

定患者有此病，并且开始治疗吗？如果试验结果需要由训练有素的阅片者（如放射医生）解释，医生应该考虑这些阅片者意见吗？

为了回答上述问题，医生需要有该试验的绝对与相对性能的信息，并了解试验和训练有素阅片者间的复杂交互作用（Beam et al.，1992）。医生必须问，对于“有病”患者，该试验执行（即灵敏度）如何？对于“无病”患者，该试验执行（即特异度）又如何？该试验可取代旧试验或应该采用多项试验吗？如果做多项试验，应该如何综合这些信息（即采用串联还是并联）？不同阅片者解释的重现性如何？

影像片的质量通常会影响诊断准确度的评价，正如 Lusted（1971）所强调的，影像可以从自然的角度最忠实再现组织的形状与结构，但它不可能装载有用的诊断信息。Fryback 和 Thornbury（1991）提出了评价医学诊断试验功效的工作模型，该模型勾画了影像质量、诊断准确度、治疗决策和病人结局对诊断试验评价的影响。基于他人的工作（Cochrane，1972；Thornbury，Fryback，and Edwards，1975；McNeil and Adelstein，1976；Fineberg，1978），Fryback 和 Thornbury（1991）提出了以下 6 级层次模型：1 级位于底层，为技术功效（technical efficacy），由拍片试验的影像分辨率与清晰度、诊断标识物试验的最佳取样时间与剂量等指标反映；2 级为诊断准确度功效（diagnostic accuracy efficacy），包括灵敏度、特异度和接受者工作特征（receiver operating characteristic，ROC）曲线；3 级为诊断思维功效（diagnostic thinking efficacy），如可由医生已知试验结果前后的诊断估计概率的差值等来度量；4 级为治疗功效（therapeutic efficacy），可由诊断试验赢得的计划治疗时间百分率来度量；5 级为病人结局功效（patient outcome efficacy），如可由试验信息减少的死亡数，改善的生存质量等来度量；6 级位于顶层，为社会功效（societal efficacy），通常由试验的成本效益指标度量。该模型的主要特点是：如果诊断试验在较高级别有效，则其所有更低级别必定有效；反过来不成立。本书只探讨诊断准确度功效（层次模型的第 2 级），这只是完整评价诊断试验有效性的一个步骤。

1.3　诊断医学统计学方法的历史回顾

于 1975 年，Lusted 在 Science 杂志上发表了较有影响的文章，文中指出：为了衡量某诊断试验的价值，必须测量试验观测者的能力，ROC 曲线提供了研究观测者能力的理想工具。尽管 Lusted 的文章只评价了 X 线片观测者的诊断准确度，但 ROC 曲线目前已广泛用于其它许多医学领域。

ROC 曲线是采用诊断试验灵敏度与假阳性率［即（1－特异度）］绘制的图形。曲线反映了随灵敏度增加，试验假阳性率的改变情况。

ROC 曲线及其分析以统计决策理论为基础，起源于电子信号观测理论（Peterson et al.，1954；Swets et al.，1982）。已用于许多医学、非医学领域，如人类感知和决策研究（Green 等，1966）、工业质量控制（Drury et al.，1975）、军事监控（Swets，1977）等。

正如电子信号观测，诊断医学中的观测者可用 ROC 曲线判断诊断试验有无区分疾病的能力。Lusted 给出的例子如下：图 1.1 中的 6 个点分别代表 6 名医生的诊断，假

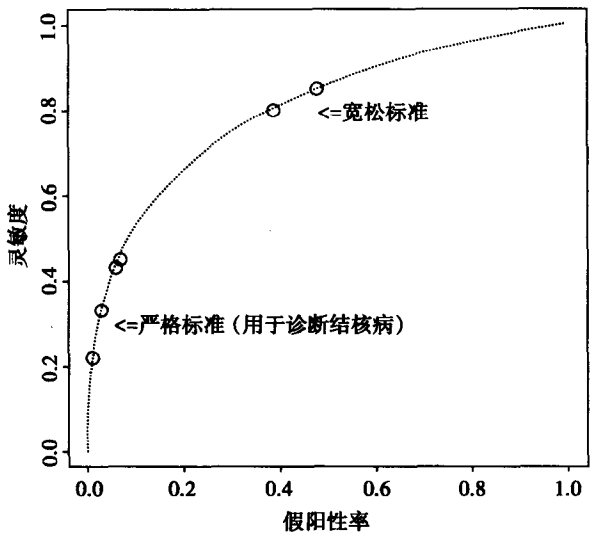


图 1.1 Lusted（1971）的结核病观测实例

定医生们具有相同的感觉和感知能力，但具有不同的诊断结核病的标准。医生根据同一张胸片区分患者有无结核病，曲线上方的点代表医生具有更宽松的决策标准（即低密度结点为阳性），而曲线下方的点代表医生具有更严格的决策标准（即仅仅高密度结点为阳性）。诊断医学中关注的是观测者解释试验结果的能力，而不是他的决策标准。

Swets 和 Pickett（1982）指出了 ROC 曲线的其它两个重要特征：第一，曲线显示了所有可能的诊断界点，每个界点提供了不同结局（即真阳性、真阴性、假阳性、假阴性）的频数估计（定义见第 2 章）；第二，曲线允许利用以前的疾病发生概率、以及正确与不正确决策的成本效益估计值，在特定情况下确定某试验的最佳诊断界点（见第 2 和第 4 章）。

Green 和 Swets（1966）首次采用正态模型估计 ROC 曲线，他们假定各种感官事件（即试验结果）可以绘制成一条简单直线。观测事件数字值（记为 T）影响了观测者关于有无疾病的信念，如果 $T<c$（c 为界点值），则观测者选择“无病”；如果 $T>c$，则观测者选择“有病”。此外，对于每种情况，假设 T 服从正态分布，基于这些假定，Dorfman 和 Alf，Jr.（1968，1969）提出了双正态（即两个通常重叠的正态分布）ROC 曲线参数的最大似然估计（MLEs）以及获得其方差协方差矩阵和置信区间的方法（见第 4 章）。为了执行 MLE，他们还编写了 FORTRAN 程序（RSCORE）。

10 余年后，Metz（1978）以及 Swets 和 Pickett（1982）根据实际情况描述了 ROC 曲线研究的设计以及分析这类数据的方法，特别强调采用 ROC 曲线下面积作为试验准确度的指标。Metz 及其同事修改和扩展了 MLE 软件 RSCORE，他们采用 FORTRAN 语言编写的程序有 ROCFIT、LABROC、CORROC 和 CLABROC，目前仍被广泛用来估计和比较双正态模型 ROC 曲线的参数。

Hanley 和 McNeil 于 1982 年给出了无任何试验结果分布假定的情况下，简单计算 ROC 曲线下面积的方法。文中指出：ROC 曲线下面积值与 Wilcoxon 两样本检验统计

量等价，这一令人感兴趣的问题首次由 Bamber（1975）描述。这种等价性使得 ROC 曲线下面积的解释更加简单，目前正被广泛应用。ROC 面积研究的样本含量估计方法是 Hanley 和 McNeil（1982）的另一个重要贡献。自此以后，估计和比较 ROC 曲线的其它非参数方法（见第 4 和第 5 章）、样本含量的估计方法（见第 6 章）被相继提出。

Swets 和 Pickett（1982）首次分析处理了多阅片者研究数据，这类数据由多个观测者解释相同患者的试验结果。他们指出了多阅片者研究中的变异和相关的几个来源，提出通过计算不同的变异成分和相关，来估计和比较多阅片者研究的试验准确度。目前可采用好几种方法来分析多阅片者研究数据（见第 9 章）。

针对有序数据，Tosteson 和 Begg（1988）首次提出了采用一般回归模型估计 ROC 曲线的方法，用这些模型可考虑协变量（如病人的年龄、性别）对试验准确度的影响。自此以后，新的回归方法以及基本模型的扩展方法被相继提出（见第 8 和 9 章）。

因为 ROC 曲线下面积包括了所有的假阳性率（从 0.0 到 1.0），是一个试验准确度全局指标，所以 McClish 于 1989 年提出了估计和比较部分 ROC 曲线下面积的参数方法，这些方法以双正态模型为基础，与完整 ROC 曲线下面积的 MLE 方法（见第 2、第 4 和第 5 章）类似。

Ransohoff 和 Feinstein（1978）对诊断试验研究设计进行了研究。他们指出，在诊断试验灵敏度、特异度估计中可能出现两个普遍问题：第一，除非广泛选取有与无疾病的对象，否则研究可能获得过高估计的灵敏度与特异度，此称频谱偏倚（spectrum bias）；第二，除非试验的解释与真实诊断的确立是独立完成的，否则偏倚可能错误地提高试验的准确度估计值，此称全面检查偏倚（workup bias）。他们采用几个诊断试验实例阐明了这些问题。自此以后，诊断试验准确度研究的其它问题也被相继发现（见第 3 章）。

在此研究后不久，校正偏倚数据的许多统计学方法被提出。如 1980 年 Hui 和 Walter 提出了标准试验存在错误［即存在不完善金标准偏倚（imperfect gold standard bias）］时，估计诊断试验灵敏度和特异度的方法；1983 年 Begg 和 Greenes 提出了消除证实偏倚（verification bias）对灵敏度和特异度的影响的方法。自此以后，许多其它解决不完善金标准偏倚（见第 11 章）和证实偏倚（见第 10 章）的方法被相继提出。

最近，综合诊断试验准确度研究（即 Meta 分析）的统计学方法有了较大进展。在所有研究无相同诊断界点的假定（通常无效，见第 7 章）条件下，Littenberg，Moses 和 Rabinowitz（1990）建议采用综合接受者工作特征（summary receiver operating characteristic，SROC）曲线作为总结试验灵敏度和特异度的工具。自此以后，又有许多基于 SROC 曲线的新方法被提出（见第 12 章）。

1.4　软件

本书讨论的许多统计学方法的计算程序可以免费获得。其中一部分是 FORTRAN 格式程序；另一部分是 SAS 宏程序（SAS Institute，Cary，North Carolina，USA）。作者已准备一个网址，容纳、链接或引用与诊断医学统计学方法相关的软件，这个网址

是：http://faculty.washington.edu/~azhou/diagnostic.html，在本书出版后至少5年内将定期维护与更新其内容。

1.5 本书没有包含的主题

尽管本书覆盖了诊断医学统计学方法的绝大多数内容，但还有下列内容没有包括。

本书讨论了运用ROC曲线描述与比较诊断试验准确度的方法。ROC曲线、特别是ROC面积也被用来评价拟合模型的预报能力。例如，SAS的LOGISTIC过程输出的c统计量与ROC曲线下面积的非参数估计结果等价，在LOGISTIC过程中用来描述拟合模型区分两组能力的大小。有关这一特殊应用ROC曲线有许多参考文献（Harrell, Jr.，Lee，and Mark，1996；Hosmer and Lemeshow，2000）。

决策分析、成本效果分析和成本效益分析，通常用来定量某试验对患者与社会引起的长期效应。本书的第2和第4章讨论了如何采用这些方法寻找ROC曲线的最佳界点，但如何执行这些方法的描述，超出了本书范围。关于这一论题，有许多优秀文献可供参考（Pauker and Kassirer，1975；Weinstein et al.，1980，1996；Russell et al.，1996；Gold et al.，1996）。

许多诊断试验可用于无症状患者的筛查，或用于已知病情患者的监控，本书描述的许多统计方法也适用于这些试验，但本书没有涉及这些试验应用时存在的具体问题，对于这些问题可参见Morrison（1992）、Murtaugh（1995）、Black和Welch（1997）。

本书给出的统计学方法大多数为频率学派观点。也可采用贝叶斯方法，将以往获得的信息、试验特征的专家观点、患者或人群的信息应用到诊断试验评价中。贝叶斯方法用于诊断试验评价的文献有Hellmich等（1988），Gatsonis（1995），Joseph、Gyorkos和Coupal（1995），Peng和Hall（1996），O'Malley等（2001），等等。

本书给出的方法适用于只有两个疾病状态（如帕金森病有与无）的情况，但在某些情况下，有多于两个真实疾病状态（如胸片的结果有气胸、间质病、结节或正常）的情况。评价多个真实疾病状态的诊断试验准确度可参考Steinbach和Richter（1987），Rockette（1994），Mossman（1999），Obuchowski、Lieber和Powell（2001）等文献。

本书没有讨论诊断试验评价的规范性要求，这些要求可以在相应管理机构维护的网站中找到。

当一个人接受多项诊断试验时，为了综合这些试验的信息，作出最佳诊断，可参见Pepe和Thompson（2000）的文章。

1.6 小结

医生需要懂得如何选取和解释诊断试验，但关于诊断试验评价的大多数文献质量较差，由此导致了误解和不可信。大量研究探讨了诊断试验准确度的设计、分析和解释方法，本书对这些方法进行了全面的综合与解析。

诊断试验评价的统计学方法不断地得到发展、修正和扩展，和读者一样，作者期盼